Week 2 Report

The dataset under analysis contains a total of 69 columns, out of which two columns are numerical, and the rest are categorical. The purpose of this report is to provide an overview of the data, identify any existing problems, and describe the approaches used to address these issues.

During understanding the dataset, the following problems were encountered:

Missing Values:

There are columns in the dataset that contain missing values. Notably, the columns "Change_Risk_Segment," "Change_T_Score," "Tscore_Bucket_During_Rx," and "Risk_Segment_During_Rx" exhibit a substantial number of missing entries. It is important to address these missing values to ensure the integrity and accuracy of the analysis.

Outliers:

The column "Dexa_Freq_During_Rx" contains a significant number of outliers, totalling 460 observations. Outliers can significantly impact statistical analyses and modeling outcomes. Therefore, it is crucial to identify and appropriately handle these outliers.

Imbalance in Target Feature:

The target feature in the dataset shows an imbalance, with a majority of observations falling into the non-persistent category. This class imbalance may introduce biases and affect the performance of predictive models.

Approaches to Address the Issues:

To overcome these problems, the following approaches are being considered:

Missing Values:

Various imputation methods will be employed to handle the missing values in the dataset. These methods include mode imputation or advanced techniques such as regression imputation.

Outliers:

There are 460 records which considered outliers in the Dexa_Freq_During_Rx column and planning to remove those as they are significantly less amount.

Sensitivity Analysis:

Due to the significant number of missing values in the Change_Risk_Segment column, a sensitivity analysis will be conducted. This analysis will involve imputing the missing values using different methods and assessing the impact on the results. Sensitivity analysis helps in understanding the robustness of the analysis to missing data.

Exploratory Data Analysis (EDA):

EDA will be conducted to gain insights into the relationships, patterns, and distributions of the data. This will involve visualizations such as histograms, scatter plots, and box plots to understand the data distribution, identify potential correlations, and uncover any additional data issues that need to be addressed.